DOCUMENT RESUME

ED 310 115 TM 013 557

AUTHOR Enright, Gwyn
TITLE Evaluation, Testing and Learning Assistance.
PUB DATE Jun 88
NOTE 20p.; Paper presented at the Annual Institute for Learning Assistance Professionals (Long Beach, CA, June 1988).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Computer Assisted Testing; *Educational Testing; Educational Trends; *Evaluation Utilization; Higher Education; Instructional Effectiveness; Student Evaluation; *Teacher Role; Test Use
IDENTIFIERS Fairness; *Learning Assistance

ABSTRACT

The resurgence of testing and the significance of this trend to learning assistance facilitators and developmental educators are discussed. This increase in testing, particularly as it manifests itself on college and university campuses, reflects a shift of emphasis in testing, from measuring aptitude to measuring the effect of instruction. Various direct and indirect forms of measurement are discussed; and new testing modes, including computer-assisted testing, are covered. The relation between the test taker and the test and the use of tests and other forms of assessment are briefly outlined. Finally, evaluating test design for lucidity, meaningfulness, economy, and fairness is promoted; and the role of the learning assistance facilitators and developmental educators in the evaluation of testing is discussed. (TJH)

Evaluation, Testing and Learning Assistance

Gwyn Enright

English Department

San Diego City College

Paper presented at the Annual Institute for Learning Assistance Professionals, California State University, Long Beach, CA, June, 1988

Evaluation, Testing and Learning Assistance
Abstract

Testing is enjoying a rebirth in educational circles, and learning assistance professionals and developmental educators are uniquely qualified to become involved. Testing is not, in and of itself, unfair. To evaluate tests, one needs to judge the tests themselves, to cultivate a healthy skeptism, and to keep current on new developments in testing. To the knowledgeable professional, tests and assessment can contribute to a better program.

EVALUATION, TESTING AND LEARNING ASSISTANCE

Recently, testing has enjoyed renewed activity on college and university campuses. This isn't the first time educators have promoted testing. The historian Daniel Resnick traces parallel developments through two periods of intense interest in testing. Both are typified by efforts to maintain quality in a growth period. The first, from 1900 to 1925, was a time of growth in secondary education. The second, from 1960 to the present is a period of growth in higher education. Both show testing's importance for public accountability when enrollments are increasing rapidly.

New Developments

Yet, Resnick points out differences in the most recent resurgence of testing. For one thing, he claims the true believers in testing are gone; today, we are more realistic about assessment. He attributes this change to a change in emphasis from testing the endowments, or aptitudes, of individuals to testing the effects of instruction on individuals. Testing is no longer a magical event that reveals the truth about students' inherent and unchangeable make-up. Now assessment systems are like maps to discern as closely as possible where a student has been in his or her educational journey as well as what he or she has taken and kept from those places. Information about learners is taken from more than one data source. Testers triangulate, as Adelman

(1986) from the Office of Educational Research points out, to make the best possible decision given the odds.

Testing is no longer just a sorting process. In many cases, we are not interested in spreading scores out. For example, to show content mastery at the end of an instructional program, criterion referenced or domain referenced testing might be more appropriate. However, commercially available tests called CRTs are not necessarily CRTs. Criterion referenced testing has more to do with the way the test is developed than any cut off or "criterion" used in scoring it (Popham, 1975).

Now, assessment does not have to be limited to objective tests. Comparable reading measures might include written retellings, which give more insight into the comprehension process itself than conventional reading tests (Smith and Jackson, 1985). New ways to measure reading reported over the last four years in the Journal of Reading, the Journal of Developmental Education, the National Reading Conference Yearbook, and the Journal of College Reading and Learning include informal reading inventories, retellings and summaries, and reading protocols.

Alternatives to objective tests bring with them another development: the debate over direct versus indirect measurement. Proponents of direct assessment argue that the meaningfulness of the construct being measured outweighs the costs of the test.

Carnegie foundation president Boyer warns of the danger of "measuring by yardsticks that matter least" (Jacobson, 1986). Guthrie (1984) points out that what we value is difficult to assess, and he warns that our renewed interest in testing must be accompanied by better measures than in the past. He writes, "It is utterly implausible that multiple choice questions could be formulated to do justice to higher skill levels," and he concludes, "Although free response formats pose serious challenges of time, logistics and reliability, the need for exploring and developing them is directly proportional to the pervasiveness of testing as a dimension of schooling" (p.190).

A corollary to the debate of direct versus indirect testing is the question of how subjectively the tests are scored. This, too, is a validity issue since notions of quality are not stable across individuals nor historical periods (Witte, Trachsel, & Walters, 1986; Stedman & Kaestle, 1987; Gould, 1981). For example in writing assessment, the assumptions on which scoring rubrics are based and the assumptions being made in designing the writing tasks are based on specific ideas of what it means to be literate. These assumptions may vary from writing assessment to writing assessment. However, on the other hand, controlling the subjectivity of direct testing formats can result in an assessment which looks like a direct assessment, but which like an indirect assessment really examines only the minute parts of a skill (Langer and Applebee, 1987). Thus the advantages of direct

assessment are lost.

Besides new testing formats, there are also new testing modes. The Computer Placement Test from the College Board uses items from the New Jersey Basic Skills Assessment. It is adaptive testing; all students need not answer all questions. A range is determined according to the students' responses. The Educational Testing Service (ETS) has pilot tested a version of computerized testing, called Computerized Mastery Testing, with parts of the Architect Registration Examination - the board exam administered to prospective architects. The efficiency of the computer mode allows more accurate testing with fewer test items (ETS Developments, 1988). An authoring system for adaptive testing (MicroCAT) is available and has been used by public schools, the U.S. Navy, and the University of Illinois (Feuer, 1986). While students may perform differently when they take a computer version of the test compared to when they take the paper and pencil version (Heppner et al, 1985), the computer mode of administering tests definately merits further investigation.

The interaction between the test taker and the test has commanded attention recently. Differential test performance, especially for developmental students, may result from the speededness (Kerstiens, 1986; Flores & Seaman, 1978) or from the modality (Enright, 1988) of the test. In addition, the test taker's race (Flores & Seaman, 1978; Immerman, 1980; White & Thomas, 1981) or

second language proficiency (Alderman, 1981) may be associated with test scores. The interest in these interactions is part of the recent effort to disentangle exactly what it is tests measure.

Finally, the way in which we regard the business of testing is changing. Aptitude, once considered a stable trait which predicted academic success, is a misnomer according to Resnick. The Scholastic Aptitude Test (SAT) measures "schooled abilities" and "aptitude in testing" instead of potential for learning. Owen, in his book None of the Above, describes testing as a cult. He ridicules the hysteria that accompanied the decline in this country from 1960 to 1980 of SAT scores:

> There is no better illustration of our national test mania than the response of otherwise rational people to the decline in SAT scores that began in the early 1960s and apparently ended in the early 1980s. As the average scores dipped year after year, dozens of explanations were advanced: nuclear fallout, junk food, cigarette smoking by pregnant mothers, weather ("Every state with an average of the math and verbal SAT scores of 510 or above also had an average high temperature in January of less than 42 degrees," according to Psychology Today.), food preservatives, declining church attendance, television ("Is television a cause of the SAT score decline? asked a College Board advisory panel in 1977, answering without hesitation, "Yes, we think it is."), the military draft, the assassination of President Kennedy, the increased incidence of marriage among female teachers, communism, pornography, the NAACP, the reduced number of eldest children in the testing population, teenage alcoholism, the American Civil Liberties Union, fluoridated water, women's liberation, witches, the civil rights movement, the war in Vietnam, the increased use of anesthesia in childbirth, and hippies (p.10).

The upturn of SAT scores brought sighs of relief. In "Student

Change, Program Change: Why SAT Scores Kept Falling," ETS scholar Turnbull wrote that the declining scores in the 1970's were a result of accommodating the needs of less proficient students in the school program to the detriment of average and high achieving students ("The Declining Score Mystery Solved?" 1986). Ironically, this reflective, post crisis stance is assumed in the face of relatively small SAT gains and is destined to be only temporary. When 1986 SAT scores failed to climb higher than 1985 scores, the press quoted former Education Secretary Bell. Bell called the failure, "...the worst news we've had in a long time" (Bell, 1986).

Owen (1985, p.10) goes on to cite with relish a 1967 study in which University of Michigan researchers found a blood test a better predictor of student completion rate than the SAT. Owen's position is the Educational Testing Service holds an arrogance operationalized in the cloaking of its exams and in its attitude that testing is not public business. In order to maintain this attitude, according to Owen, ETS plays on most people's insecurities about tests.

Using Tests and Assessment

However, learning assistance professionals and developmental educators should not be intimidated by tests. Longtime proponents of program evaluation (Boylan, 1984; Devirian, 1973; Maxwell, 1979; Walvekar, 1987), learning assistance professionals

and developmental educators may be uniquely qualified to participate, and perhaps to lead, in the testing arena. They established evaluation as a crucial component of learning assistance support systems when other, more entrenched campus departments and agencies ignored or resisted serious program evaluation. For twenty years, learning assistance professionals have taken a proactive stance vis a vis evaluation. As a result, their experience included developing, refining, using and interpreting evaluation instruments. And these skills apply to selecting and guiding the use of tests.

Uniquely qualified, learning assistance professionals and developmental educators will benefit from participating in the new testing movement. Assessment contributes to solid programs. Suanne Roueche (1983) reports a national study in which she identifies elements of successful programs. She defines success as 50% or better student retention in basic skills classes. Of the eleven elements identified, at least two pivot on a strong assessment program: 1) mandatory counseling/ placement and 2) program evaluation. Obler (1983), in her article, "Programs for the Underprepared Student: Areas of Concern,' stresses assessment as the <u>first</u> area of concern.

Assessment seems to be fundamental and imperative. It should be as thorough as possible. Testing is still fraught with controversy. However, we have grown from civil rights paranoia to concern for congruence between mission and practice. The more we know, the better we can serve students. It can even be argued that it is immoral not to test students who might otherwise attempt advanced work, thereby committing "academic suicide." (p.

In addition, Forrest (1982) states the single most important element contributing to student retention is orientation or guidance at the beginning of a student's career. Orientation includes course placement based on tests and academic records.

Direct involvement of learning assistance professionals and developmental educators in testing issues should enhance the effectiveness of assessment systems being put into place. In Clowe's interview with Richardson, Richardson blames the missing connection between the instructional outcomes of developmental courses and the entry requirements of advanced courses for a Maryland community college's finding that the best preparation for college-level English composition was to take the course once and fail it. "Students who followed that course of action did better than those who successfully completed the developmental courses in English" (Clowes, 1986). An assessment system in which English instructors and developmental educators specify outcomes and select appropriate measures might rectify the indefensible situation Richardson reported.

Evaluating Tests

Even though learning assistance professionals and developmental educators have developed and used evaluation instruments such as student surveys and faculty questionnaires, these same practitioners tend to be suspicious of tests per se. Yet, tests

can be fair if they are designed and used thoughtfully. When the director of the Center for the Study of Evaluation, Baker, spoke at a 1978 conference on Measurement and Methodology in Education, she asked, with regard to tests, "Is Something Better than Nothing?" Her answer was, "Yes, if." The "if," of course, was the clincher. Baker went on to list at least three conditions. Her conditions included deriving lucid test specifications in order to make test information public without publishing every test item and in order to guard against arbitrarily generated test items. Another condition was meaningfulness. In other words, tests ought to have merit and value in and of themselves. The rationale supporting this condition is the power tests have for students as well as the power tests have to shape the curriculum. A final condition was economy. Only test data that will be used should be collected. The sum total of Baker's "if" is fair testing and fair testing "implies open, public, meaningful, coherent, and economic testing practices" (p.30).

Besides Baker's conditions for test design, Maxwell (1979) noted a programmatic consideration for viable testing practices. Testing generates the concomitant need for additional services. It doesn't make sense to test learners if the program doesn't exist to meet the needs identified. So, the requisite program is a requirement for fair testing. These are all non trivial conditions to be met in order to endorse testing.

There are requirements needed for evaluating tests. First, time is needed to read all test items and determine whether they seem to measure specific skills they purport to measure or those skills one is interested in testing Relying on the categories of what specific test items are measuring that are provided by the test manual instead of making this determination oneself may lead to errors in matching tests to instruction. Next, one needs the energy to pester the test publisher into sending the test itself, the directions for administering the test, the index of instructional objectives and the technical manual that should accompany the test. The specimen set alone is not enough. Finally, and most importantly, one needs a healthy skepticism. This last requirement for evaluating tests is especially handy when reviewing test manual claims, for publishers are inclined to omit data that detracts from their test and highlight data that exalts their product (Gordon, 1983; Kerstiens, 1986).

By evaluating tests or examining tests carefully, learning assistance professionals and developmental educators can help ensure testing is fair. For example, in a paper published in the Journal of College Reading and Learning, Condon (1987) described the ten step exam analysis she uses to help nursing students specify exactly why they missed certain test questions and to prescribe exact courses of action they can take to improve their performance on future exams. Condon's analysis of individual test items helps demystify tests and also contributes

to the meaningfulness and economy of the test.

Wood (1987) carefully examines four tests. She discusses the advantages and disadvantages of the Nelson-Denny, the Degrees of Reading Power, the Descriptive Test of Language Skills, and the New Jersey College Basic Skills Placement Test. She concludes in a sober tone that current standardized reading tests can be used only as gross measures of reading ability. "They test such a limited part of what most of us teach in reading classes that when we give them as post tests, we must remind ourselves, our administrators and our students that they by no means measure all that students have been taught or are able to do. In fact, if we limit our teaching to improving the scores on these tests, we will not be providing our students with the skills and abilities they need to perform college reading tasks effectively" (p.12).

Wood's evaluation of four tests emphasizes the need for meaningfulness in testing; her warning is based on the tendency to allow tests to drive the curriculum. Moving toward fair testing is the reward for evaluating the tests themselves by examining them closely, test item by test item.

The Challenge: Taking an Active Part

Assessment is enjoying a renaissance, and learning assistance facilitators and developmental educators must take an active role

in student assessment issues. Learning assistance professionals and developmental educators must take an active role in participating on any campus committee to select or to evaluate screening instruments used or to be used at the college; learn about new developments and new approaches in diagnostic and achievement measures in developmental education; read test reviews and research reports; demand the technical manuals that accompany or should accompany tests; and maintain an appropriate skepticism.

In taking this challenge to become actively involved in testing, learning assistance professionals and developmental educators will contribute toward fair testing practices. Popham makes a nifty distinction; he distinguishes between educators and measurers (1978). We need more educators in the assessment arena. To quote the Ralph Nader of standardized testing, "Testing is too important to be left to the test makers" (Owen, 1985). And testing is also too important to be left to other decision makers on campus who may be less sensitive to the importance of tests than learning assistance professionals.

References

Adelman, C. (1986, February) Discussant at Testing and quality assurance in higher education: The role of entry, classroom and exit examinations, conference sponsored by Miami-Dade Community College Office of Institutional Research, Miami, FL.

Alderman, D. L. (1981). Language proficiency as a moderator variable in testing academic aptitude. Unpublished manuscript (RR 81-41), Princeton: Educational Testing Service.

Baker, E. L. & Quellmalz, E. S. (1980). Educational testing and evaluation: Design, analysis and policy. Beverly Hills: Sage Publications.

Bell calls SATs bad news. (1986, June). Los Angeles Times, 2.

Boylan, H. R. (1984, February). Evaluating remedial programs: The state of the art and beyond. Paper presented at the invitational conference on the evaluation of remedial programs in post secondary education, Asilomar, Ca.

Breakthrough development in computerized testing offers shorter tests, more precise pass-fail decisions. (1988, Winter/Spring). ETS Developments, 3 & 4. Princeton: Educational Testing Service, 3-4.

Clowes. D. (1986). Literacy in the open-access college: An

interview with Dr. R. C. Richardson, Jr. Journal of Developmental Education, 10(1), 16-21.

Condon, V. (1987). The exam analysis. In J. L. Mullen (Ed.) Journal of College Reading and Learning, 147-154.

The Declining Score Mystery Solved? (1986). Journal of Developmental Education, 10(1), 28-29.

Devirian, M. C. (1973). Data Collection: A Cybernetic Aspect of a Learning Assistance Center. In G. Kerstiens (Ed.) Proceedings of the Sixth Annual Conference of the Western College Reading and Learning Association, Albuquerque, N.M., 51-58.

Enright, G. (1988, May). Measuring college reading and writing skills. Paper presented at the Thirty-third Annual Conference of the International Reading Association, Toronto, Canada.

Feuer, D. (1986, May). Computerized testing: A revolution in the making. Training, 80-86.

Flores, T. B. & Seaman, D. F. (1978). A comparative study of adult student performance on timed GED tests in Texas. (ERIC Document Reproduction Service No. ED 154 138)

Forrest, A. (1982). Increasing student competence and persistence: The best case for general education. A report of the College Outcome Measures Project (COMP). Iowa City: American College Testing Program.

Gordon, B. (1983). A guide to postsecondary tests. Reading World, 23, 45-53.

Gould, S. J. (1981). The mismeasure of man. New York: W.W. Norton.

Guthrie, J. T. (1984). Research: Testing higher level skills. Journal of Reading, 28, 188-190.

Heppner, F. H., Anderson, J. G. T., Farstrup, A. E. and Weiderman, N. H. (1985). Reading performance on a standardized test is better from print than from computer display. Journal of Reading, 28, 321-325.

Immerman, M. A. (1980). The effect of eliminating time restraints on a standardized test with American Indian adults. (ERIC Document Reproduction Service Report No. ED 196 484)

Jacobson, R. L. (1986, October 15). Efforts to assess students' learning may trivialize the B. A., Boyer says. The Chronicle of Higher Education.

Kerstiens, G. (1986). A testimonial on timed testing. In M P. Douglass (Ed.) Fiftieth Yearbook of the Claremont Reading Conference. Claremont, Ca: Claremont Graduate School, 261 - 268.

Kerstiens, G. (1986, June). Post-secondary reading comprehension testing: The late debate. Unpublished manuscript, Andragogy Associates.

Langer, J. A. & Applebee, A. N. (1987). How writing shapes thinking A study of teaching and learning. (NCTE Research Report No. 22). Urbana: NCTE, 146-148.

Maxwell, M. (1979). Improving student learning skills. San

Fransisco: Jossey-Bass.

Obler, S. S. (1983). Programs for underprepared students: Areas of concern. In J. E. Roueche (Ed.), A new look at successful programs. New Directions for College Learning Assistance, San Fransisco: Jossey-Bass, 11, 21-31.

Owen, D. (1985). None of the above: Behind the myth of scholastic aptitude. Boston: Houghton-Mifflin.

Popham, J. (1978). Modern measurement technology: Miracle or mirage? In G. Enright (Ed.) Proceedings of the Eleventh Annual Conference of the Western College Reading Association, Long Beach, CA, 12-15.

Popham, J. (1975). Education evaluation. Englewood Cliffs: Prentice-Hall.

Resnick, D. (1986, February). Testing and examinations in historical perspective. Keynote address at Testing and quality assurance in higher education: The role of entry, classroom and exit examinations, conference sponsored by Miami-Dade Community College Office of Institutional Research, Miami, FL.

Roueche, S. (1983). Elements of program success: Report of a national study. In J. E. Roueche (Ed.), A new look at successful programs. New Directions for College Learning Assistance, San Fransisco: Jossey-Bass, 11, 3-10.

Smith, S. P. & Jackson, J. H. (1985). Assessing reading/learning skills with written retellings. Journal of Reading, 28, 622-630.

Stedman, L. C. & Kaestle, C. F. (1987, Winter). Literacy and reading. Reading Research Quarterly, 1, 8-46.

Witte, S. P., Trachsel, M. & Walters, K. (1986). Literacy and the direct assessment of writing: A diachronic perspective. In K. L. Greenberg, H. S. Wiener, and R. A. Donovan, (Eds.) Writing assessment Issues and strategies (pp. 13 - 34). N.Y.: Longman.

Walvekar, C. C. (1987, April). Thirty years of program evaluation: Past, present, future. In J.L. Mullen (Ed.) Journal of College Reading and Learning, 155-161.

White, E. & Thomas, L. (1981). Racial minorities and writing skills assessment in the California State University and Colleges. College English, 42, 276-283.

Wood, N. V. (1987, April). Standardized reading tests and the post secondary reading curriculum. Paper presented at the 20th Annual Conference of the Western College Reading and Learning Association, Albuquerque, N.M.